Evaluating the Impact of Retrieval-Augmented Generation on Conversational Agents for Medical Question Answering

Bohan Yu and **Guanyu Yang** and **Hamza Mooraj**
and **Jamie Lawson** and **Xinyu Zhu** and **Yang Yang**
and **Yu Zhou** and **Yuqi Wang** and **Zhaoyang Liu**
Heriot-Watt University
{by2006, gy2004, hhm2000, jl2132, xz2025, yy2039, yz2069, yw2035, zl2045}@hw.ac.uk

Abstract

Medical question answering (QA) systems face challenges with hallucinations, limiting their utility in healthcare. This study explores the impact of Retrieval-Augmented Generation (RAG) in improving the accuracy of such systems. A tri-modal conversational agent (CA) architecture is proposed, combining text, speech, and visual inputs, leveraging a large language model (LLM), Whisper for speech recognition, and Qwen2-VL for visual processing. The LLM is fine-tuned on the MedQA dataset, and Qwen2-VL on PMC-VQA. FAISS is used for efficient retrieval of information from the NHS Inform Scotland dataset. The system's performance is evaluated before and after applying RAG. Results from intrinsic evaluation demonstrate that RAG enhances factual accuracy and semantic similarity. However, heuristic evaluation reveals usability issues, including complex medical terminology and inconsistencies in multimodal integration. This study indicates that RAG improves accuracy in medical CAs, but usability challenges need to be addressed for effective real-world deployment.

1 Introduction

1.1 Background and Motivation

Conversational agents powered by artificial intelligence (AI) have gained prominence within healthcare, particularly in the medical question-and-answer (Q&A) domain, as further discussed in section 2.1. These agents have shown promise in improving patient interactions by delivering immediate, accurate, and personalised responses, thus facilitating enhanced clinical decision-making and patient care outcomes. However, the effectiveness of medical CAs is often limited by the issue of hallucination instances where the agent generates inaccurate or unsupported information (Aljamaan et al., 2024). Given the critical implications in healthcare contexts, mitigating these inaccuracies is crucial to ensure reliability and patient safety. RAG offers a promising approach to address this challenge by grounding the agent's responses in reliable knowledge sources. Consequently, there is strong motivation to improve the factual accuracy and dependability of CAs through this methodology.

1.2 Research Question

The primary research question guiding this project is: *How can Retrieval-Augmented Generation (RAG) enhance the accuracy and reliability of CAs for medical QA?* The research aims to evaluate whether integrating RAG techniques can effectively reduce inaccuracies and hallucinations, thereby improving the reliability and trustworthiness of medical conversational AI systems.

1.3 Project Aims and Objectives

This project aims to design, implement, and evaluate a multimodal CA tailored for medical Q&A applications. The key objectives include integrating RAG to provide accurate, contextually relevant responses, fine-tuning the Llama-3.1-8B (Grattafiori et al., 2024) language model specifically for medical domains using established datasets such as MedQA (Jin et al., 2020) (text-based) and PMC-VQA (Zhang et al., 2024)(visual-based), and incorporating multimodal input capabilities through speech recognition (Whisper) and visual understanding (CLIP). The efficacy of the RAG approach in reducing hallucinations and improving system accuracy will be thoroughly assessed, providing valuable information on the practical application and effectiveness of multimodal AI techniques in healthcare.

2 Related Work

2.1 Conversational Agents in Healthcare

CAs have seen a surge in application within healthcare, offering potential for improved patient support and information dissemination. Comprehensive reviews have mapped the landscape of these applications, highlighting their evolution and diverse roles (Valizadeh and Parde, 2022; Laranjo et al., 2018; Tudor Car et al., 2020). Early systems, while limited, laid the groundwork for more sophisticated models, as evidenced by the development and evaluation of tools like Woebot for cognitive behavioural therapy (Fitzpatrick et al., 2017). Studies have also demonstrated the feasibility of CAs in addressing a range of health-related queries via smartphone platforms, indicating their accessibility and potential for widespread adoption (Miner et al., 2016). Recent advancements, particularly in large language models (LLMs), have led to the creation of specialised models like BioGPT for biomedical text generation and ChatDoctor, fine-tuned for medical domain knowledge, showcasing the drive towards more accurate and context-aware healthcare interactions (Luo et al., 2022; Li et al., 2023). The trend towards generalist biomedical AI, as explored by Tu et al. (2023), further emphasises the expanding capabilities of CAs in handling complex medical tasks. This evolution towards more accurate and context-aware CAs, particularly with the development of specialised models like BioGPT and ChatDoctor, highlights the importance of addressing accuracy and reliability, a key focus of this study through the implementation of RAG.

2.2 Retrieval-Augmented Generation for Medical QA

The integration of RAG within medical QA has the potential to enhance accuracy and reliability, especially when dealing with dynamic medical knowledge. Recent studies have directly addressed the application of RAG in medical contexts. For instance, MKRAG introduces a medical knowledge RAG approach, emphasising the importance of integrating medical knowledge to enhance the accuracy of QA systems (Shi et al., 2024). Additionally, benchmarking studies have been conducted to evaluate the performance of RAG models in medicine, providing insights into their strengths and limitations (Xiong et al., 2024). These efforts highlight the growing recognition of RAG as a pivotal technique for grounding medical QA in reliable and up-to-date information sources. The development of models like ChatDoctor, which leverages medical domain knowledge, implicitly addresses the need for robust information retrieval to enhance response accuracy (Li et al., 2023). The use of LLMs to generate radiology reports from chest X-rays, as seen in ELIXR, highlights the importance of aligning language models with specific knowledge sources, a core component of RAG (Xu et al., 2023). This alignment ensures that generated responses are not only linguistically coherent but also grounded in reliable medical data. These studies collectively emphasise the growing importance of RAG for grounding medical QA systems with reliable information, directly motivating our research in exploring RAG's effectiveness in enhancing the accuracy and reliability of a medical CA.

2.3 Multi-modal Models and their application to Conversational Agents

Multi-modal models are increasingly recognised for their potential to enrich CAs in healthcare by incorporating diverse data types like text, speech, and images. The exploration of multimodal LLMs for health, grounded in individual-specific data, illustrates the move towards personalised healthcare solutions (Belyaeva et al., 2023). Systems like ELIXR, which combines language models with radiology vision encoders, exemplify the practical application of multimodal approaches in medical imaging (Xu et al., 2023). The increasing recognition of multi-modal models for enhancing CAs in healthcare, as demonstrated by systems like ELIXR, underpins our project's exploration of a tri-modal CA that integrates text, speech, and visual input for improved medical QA.

3 Methodology

3.1 System Overview

The CA developed in this project features two distinct models, tailored for medical QA tasks. The primary focus is a text-based model (dual-model) enhanced with RAG, which leverages the NHS Inform Scotland dataset (NHS inform Scotland, 2025) (NHS dataset) to generate accurate and contextually relevant responses. A supplementary multimodal model (tri-model), incorporating visual inputs, was also developed as a proof of concept.

The text-based model processes user queries by directly engaging with a fine-tuned language model. RAG is integrated into this process, retrieving medical information from the NHS dataset to mitigate inaccuracies and hallucinations. This integration aims to ensure the generated responses are grounded in reliable medical knowledge.

The supplementary multimodal model extends the system's capabilities to include visual inputs.

For both models, speech inputs are transcribed using Whisper and treated as textual queries.

The final outputs from both models are presented via a multimodal UI discussed in 3.6, facilitating user interaction through both textual and auditory channels. The primary evaluation focuses on the text-based model, while the multimodal model serves as a proof-of-concept for future development.

3.2 Model Selection and Fine-Tuning

The CA leverages several specialised pre-trained models selected for their demonstrated effectiveness in multimodal applications within healthcare:

Textual Model: The primary model chosen is Llama-3.1-8B (Grattafiori et al., 2024), recognised for its robust generative capabilities. This model has been fine-tuned specifically on the MedQA (Jin et al., 2020) dataset, providing a comprehensive base of medical knowledge to ensure accurate responses to user queries.

Speech Recognition Model: Whisper's base model (Radford et al., 2022) was selected for speech-to-text conversion, ensuring precise transcription of spoken inputs, which enhances accessibility and convenience for users interacting via voice.

Visual Language Model (VLM): The Qwen2-VL-2B model (Wang et al., 2024) was integrated for processing visual data, fine-tuned using the PMC-VQA dataset (Zhang et al., 2024). This allows the system to interpret medical imagery effectively, generating accurate textual descriptions that complement other modalities within the CA.

The fine-tuning procedure involved specialised datasets—MedQA for textual fine-tuning and PMC-VQA for visual understanding—ensuring these models could effectively manage the specialised terminology and nuanced medical contexts. The deliberate focus on medically relevant datasets optimises model performance specifically for healthcare applications.

3.3 Retrieval Mechanism

The retrieval mechanism is responsible for fetching relevant information from the NHS Inform Scotland medical knowledge source, which supports the generation of accurate responses by providing factual and contextually relevant information. This mechanism is implemented using FAISS (Facebook AI Similarity Search) (Douze et al., 2024) to create a high-speed, vector-based retrieval system.

3.3.1 Index Creation

The retrieval mechanism utilises a medical knowledge base sourced from NHS Inform Scotland, which provides information on various diseases, their symptoms, and recommended treatments. The dataset was created by scraping the NHS Inform Scotland Illnesses and Conditions A-Z (NHS inform Scotland, 2025) and is stored in a JSON format, with each entry containing:

- Disease: The name of the condition as described by NHS Inform Scotland.
- Symptoms: A description of symptoms associated with the disease.
- Treatments: Advised treatment options

The pre-processing steps taken to prepare the dataset for retrieval include combining each entry into a single text passage. The combined text passages are embedded into high-dimensional vector representations using the all-MiniLM-L6-v2 model from the Sentence-Transformers library. The model generates n-dimensional embeddings for each text passage in order to capture the semantic meaning.

After pre-processing, the embeddings are indexed using FAISS. The FAISS index uses the IVF-Flat indexing strategy, which was chosen for its balance of speed and accuracy, and combines IndexFlatL2 for clustering and distance calculation using L2 distance. The index is trained on the text embeddings and subsequently populated with the vectors.

3.3.2 FAISS Search

When a user submits a query, it is encoded into an embedding using the same sentence-transformer model. The query embedding is then used to perform a similarity search within the FAISS index, which retrieves the top-k most similar text passages based on L2 distance. The retrieved results include the matched text passages along with their corresponding distance scores, which represent the degree of similarity to the query.

```
 [(retrieved_text_1, distance_score_1),
(retrieved_text_2, distance_score_2),
...]
```

3.4 Pipeline for the Dual-Model

The Dual-Model pipeline is designed to process user queries provided via text or speech, utilising RAG as an optional feature. This high-level pipeline represents the working of the text-only model mentioned in 3.1.

3.4.1 Input Processing

The user input is accepted in two forms: text or speech. If the input is provided as speech, it is first transcribed into text using the Whisper model. This transcribed text is then treated as the user query for subsequent processing.

3.4.2 Response Generation

Without RAG: The user query is formatted into an alpaca-style prompt (Touvron et al., 2023), optimized for LLaMA-based models. This prompt includes an initial instruction, the user's query, and a placeholder for the model's response. The language model then generates a response using the formatted query, which is decoded using its tokenizer and post-processed to remove extraneous text.

With RAG: The user query is embedded, as detailed in 3.3.2, and used to retrieve relevant documents from the NHS knowledge source via FAISS. These documents are then concatenated into a single text string, serving as the context for the model's response. The query is re-formatted using an alpaca-style prompt, now including the retrieved context, and passed to the language model. The model is instructed to use the context to enhance response accuracy. The model's output is then decoded and post-processed.

3.4.3 Audio Output

Regardless of whether RAG is enabled, the generated text response is passed through the Google Text-to-Speech (TTS) engine. This allows the user to listen to the response as an audio output, enhancing accessibility and user experience.

3.5 Pipeline for the Tri-Model (Proof of Concept)

The Tri-Model extends the dual-model (3.4) to incorporate visual inputs alongside text or speech, serving as a proof of concept. The core functionalities for speech transcription, RAG integration for text, model generation and TTS remain consistent with the dual-model. This section focuses on the integration and processing of visual inputs.

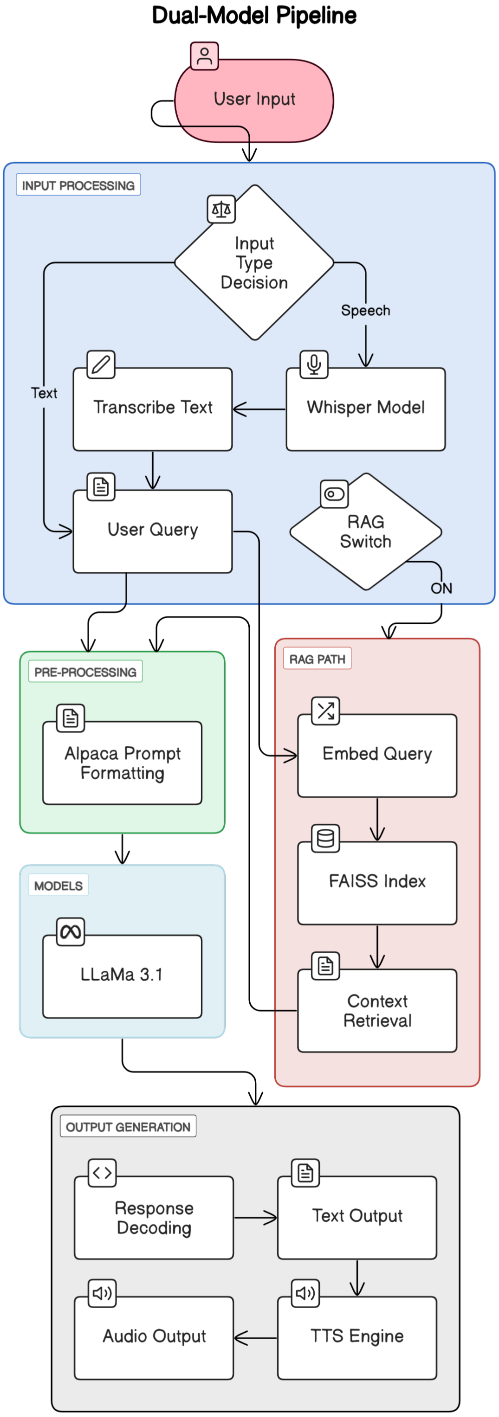


Figure 1: End-to-end inference pipeline for the Dual-Model

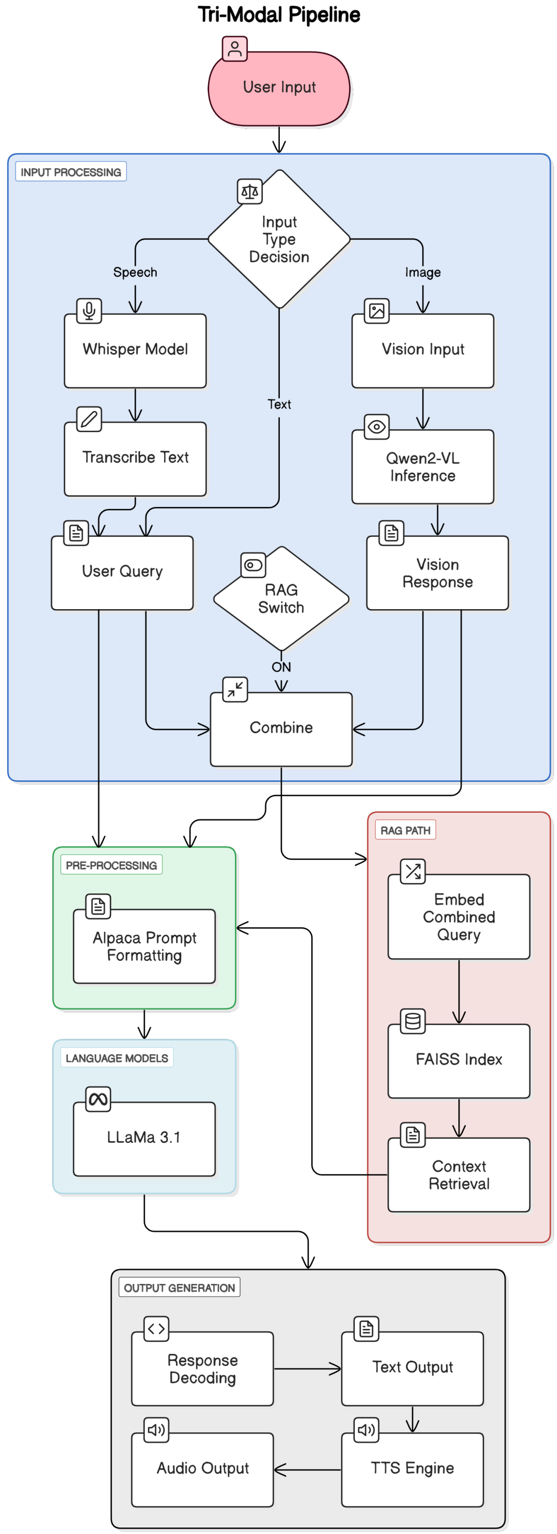


Figure 2: End-to-end inference pipeline for the Tri-Model

3.5.1 Qwen2-VL Inference

When a user provides an image, it is processed using the Qwen-VL image processor (Wolf et al., 2020). This step prepares the image for input into the Qwen2-VL model, enabling it to generate a textual description of the image and answer questions related to the image. This is the *vision response*.

3.5.2 Response Generation with Visual Input

With RAG enabled, the *vision response* is combined with the user's query to retrieve relevant documents. The retrieved information, along with the user query and vision response, is then formatted into the alpaca-prompt for the language model, which generates a response that is subsequently decoded and post-processed. When RAG is disabled, the process remains the same, except that no retrieved information is included in the prompt.

3.6 User Interface and Interaction Design

3.6.1 Overview and Interaction Modes

The system features a multimodal user interface, supporting both text and voice input. Users can type their queries or use the integrated microphone for speech input, which is then transcribed into text using the Whisper model. All inputs are processed by the back-end language model as text. The system provides output in both text and synthesized speech, using Google Text-to-Speech, enabling fully voice-based interaction.

3.6.2 Interface Implementation

Open WebUI provides the user interface, running locally via Docker alongside Ollama, which hosts the fine-tuned LLaMA 3.1-8B model in GGUF format. This setup ensures low-latency, private, and cost-effective inference. The browser-based UI includes a text entry box, voice input button, and response displays. For tri-modal usage, users can upload images for visual QA.

3.6.3 Usability and Design Considerations

The user interface is designed for clarity and accessibility, catering to users with varying technical expertise. Text-to-speech output enhances accessibility for visually impaired users, while voice input enables hands-free interaction. The on-screen conversation history aids users in tracking extensive medical interactions. Overall, the interface facilitates natural, multimodal interaction, making the CA suitable for real-world healthcare applications.

4 Evaluation

4.1 Evaluation Methods and Metrics

Intrinsic and heuristic methods were used to evaluate the impact of RAG on a medical-related CAs, with future extrinsic evaluation planned.

4.1.1 Intrinsic Evaluation

This section details how the textual quality of the model's output against the NHS dataset was measured by assessing lexical similarity, fluency, and semantic overlap between the model's response (prediction) and the NHS dataset (reference).

BLEU (Bilingual Evaluation Understudy) measures n-gram overlap between generated and reference text (Papineni et al., 2002). While common in machine translation, it's useful for short, factual answers. BLEU's use in this study quantifies surface-level similarity, giving a basic measure of correctness in the model's responses.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) evaluates recall and precision of overlapping sequences (n-grams, words, or sentences) between generated and reference responses (Lin, 2004). The following ROUGE configurations were used:

- **ROUGE-1:** Unigram overlap (single words)
- **ROUGE-2:** Bigram overlap (two consecutive words)
- **ROUGE-L:** Longest common subsequence overlap (sentence-level similarities).

Since ROUGE accounts for larger overlaps, its effectiveness for evaluating content completeness and factual accuracy coincides with the study.

BERTScore uses pre-trained contextual embeddings (from BERT) to evaluate semantic similarity (Zhang et al., 2020). It captures similarity at the meaning level, not just surface overlap. Given the semantic nature of medical language, this provides a context-aware measure of response quality.

The following evaluation strategy compared model performance with and without RAG:

- **Dual-Model:** Input queries were the 'Symptoms' feature from the NHS dataset (section 3.3.1). Model responses, with corresponding 'Disease' and 'Treatments', were stored in a CSV. This CSV provided response and reference text for metrics calculation. Evaluating responses with and without RAG directly assessed RAG's impact on accuracy against the NHS knowledge source.
- **Tri-Model:** The Tri-Model wasn't formally evaluated due to the lack of a suitable multimodal, conversational dataset. The PMC-VQA dataset, used for training, has single-word QA pairs, unsuitable for proper evaluation. Thus, the tri-model's multimodal capabilities remain a proof of concept.

4.1.2 Heuristic Evaluation

A heuristic evaluation was conducted to systematically identify usability issues within the multimodal medical CA. The evaluation specifically aimed to examine the usability aspects critical in healthcare applications, such as clarity and appropriateness of medical language, error handling and prevention, consistency across modalities, and the integration of multimodal interactions.
Adapted from Nielsen's usability heuristics (Nielsen, 2020), the evaluation criteria included:

- **Appropriateness of Language**: Assessment of whether the CA consistently uses clear and comprehensible medical terminology suitable for general user understanding.
- **Error Handling and Prevention**: Examination of the system's capability to proactively avoid inaccuracies and effectively manage errors, critical for patient safety and trust.
- **Information Consistency**: Evaluation of the consistency and reliability of the agent's responses across different interaction modes (text, speech, visual).
- **Integration of Modalities**: Analysis of how seamlessly and efficiently the CA integrates textual, auditory, and visual components to support natural user interactions.

Procedure: Internal evaluators, familiar with heuristic methods and CAs, interacted with both the dual-model and tri-model versions of the system. Evaluators conducted structured interactions, emulating typical medical inquiry scenarios. During these interactions, evaluators recorded observed usability issues according to the defined heuristics and categorised them by severity (Minor, Moderate, Severe).

4.1.3 Extrinsic Evaluation Plan

Due to undergraduate constraints, a full-scale evaluation with human participants was not conducted. The following outlines a potential future extrinsic evaluation of the developed CAs (CAs).

The **primary objective** of this evaluation is to assess the CA's performance in realistic medical scenarios, focusing on its effectiveness in assisting users with medical questions. The evaluation aims to determine the CA's impact on user task completion, efficiency, and satisfaction.

Participants: A diverse group of participants, including medical students and the general public, is proposed to reflect real-world usage. Participants will be recruited and randomly assigned to one of two groups:

- Group 1: Will use the CA **with** RAG enabled.
- Group 2: Will use the CA **without** RAG enabled.

Tasks and Scenarios: Participants in both groups will use the CA to answer realistic medical scenarios (e.g., symptom interpretation, medical information retrieval from the NHS dataset).

Metrics: The following metrics will be measured:

- **Task Completion Rates:** Proportion of participants completing scenarios.
- **Task Completion Times:** Time taken to complete scenarios.
- **User Satisfaction:** Measured via post-task questionnaire.

Questionnaires: User satisfaction will be assessed using a structured questionnaire with quantitative and qualitative questions, including:

- **Likert Scale Questions:** To assess ease of use, clarity, efficiency, and overall satisfaction (e.g., "The CA was easy to use," "The information provided by the CA was clear").

- **Open-Ended Questions:** To gather feedback on helpfulness, frustrations, and suggestions (e.g., "What did you like most?", "What could be improved?").

Procedure: The evaluation would occur in a controlled environment, with clear instructions. Data would be collected through task performance metrics and user feedback.

Quantitative data (completion rates, times, Likert scale responses) would be analysed using descriptive statistics (means, standard deviations) and comparative tests (e.g., t-tests). Qualitative data from open-ended questions would be analysed using thematic analysis.

4.2 Results

4.2.1 Intrinsic Evaluation

The intrinsic evaluation conducted produced the results seen in Table1.

Metric	without RAG	with RAG
BLEU	0.001994	0.002742
ROUGE	0.1186	0.1211
BERTScore	0.7892	0.8333

Table 1: Intrinsic Evaluation Results Comparing Performance With and Without RAG

The results indicate an improvement across all metrics when using RAG. The BLEU score, while still low, increased by **37.5%** with RAG, suggesting improvement in n-gram overlap with the reference texts. ROUGE score also saw a modest increase of **2.1%** indicating in the recall and precision of overlapping sequences. A substantial improvement was also witnessed in BERTScore between using RAG and not. An increase by **5.6%** suggests higher semantic similarity between generated responses and reference texts from the NHS dataset.

4.2.2 Heuristic Evaluation

The heuristic evaluation identified significant usability issues within the CA. Evaluators assessed responses from interactions conducted both with and without RAG. Table 2 summarises the identified issues and categorises their severity clearly.

Heuristic	Identified Issue	Severity
Appropriateness of Language	Overly complex terminology (without RAG).	Moderate
Error Handling and Prevention	Factually incorrect responses observed without RAG.	Severe
Information Consistency	Significant discrepancies between responses when RAG was toggled.	Moderate
Integration of Modalities	Latency and formatting issues with RAG enabled.	Moderate

Table 2: Summary of heuristic evaluation findings

4.3 Analysis of Results

4.3.1 Intrinsic Analysis

The intrinsic evaluation (Table 1) reveals RAG's impact on response quality. Low BLEU scores are likely due to its sensitivity to exact lexical matches. However, the increase with RAG suggests improved reproduction of knowledge source word sequences, crucial for reducing hallucinations and promoting ground-truth usage in medical CAs.

Similarly, ROUGE's reliance on exact lexical matches explains its low values. The smaller increase compared to BLEU indicates RAG has a less pronounced effect on recall of overlapping sequences, suggesting limited impact on sentence-level structure and fluency. This could stem from the model's pre-training on a globally sourced dataset (Grattafiori et al., 2024) versus the UK-specific knowledge source.

Given BLEU and ROUGE's reliance on surface-level matches, BERTScore's measurement of semantic similarity is pertinent. The notable BERTScore improvement indicates enhanced meaning-level similarity, suggesting RAG helps the model produce more accurate and contextually relevant information. This improved semantic similarity highlights RAG's benefit in enabling the model to incorporate knowledge retrieved from the knowledge source.

While BERTScore demonstrates RAG's success in semantic grounding, the smaller BLEU and ROUGE scores suggest less influence on grammatical structure and fluency. Although semantic accuracy is paramount in medical CAs, regionally appropriate language is vital for user comprehension and to prevent misunderstandings.

4.3.2 Heuristic Analysis

The heuristic evaluation dove into the CA's usability within healthcare contexts. Although enabling RAG improved the accuracy and reliability of medical information, it also introduced new usability issues. For instance, responses with RAG enabled often suffered from repetitive hyperlinks and formatting errors, potentially confusing users or obscuring essential medical information.

In contrast, when RAG was disabled, the agent occasionally generated inaccurate or irrelevant content, which had evident inaccuracies such as suggesting that sugar intake directly causes diabetes, highlighting significant limitations of non-RAG-enabled generation. Likely to undermine trust and reliability—a major concern in healthcare applications.

Findings related to language appropriateness underscored the need for clearer, simpler medical terminology, especially when explaining complex conditions like COPD and migraines. Furthermore, discrepancies between text and audio modalities indicated that multimodal integration requires further refinement to ensure coherent and consistent user experiences.

Overall, there's a trade-off between the improved accuracy provided by RAG and the associated usability challenges, pinpointing specific opportunities for future improvements in multimodal coherence, response clarity, and user-centered presentation.

5 Conclusion

Overall, this project demonstrates that integrating RAG into a multimodal medical CA enhances its factual accuracy and reliability. This section outlines the research contributions and summarises key findings from the study.

The study contributes the implementation of a RAG-enhanced multimodal CA for medical QA. The developed proof-of-concept tri-model, incorporating visual input, acts as a foundation for further experimentation in this area of research. The study also demonstrates the feasibility of using RAG to improve accuracy in medical CAs, while exploring trade-offs between accuracy and usability.

Key findings include the intrinsic evaluation showed that RAG improved performance across all metrics (section 4.1.1), indicating enhanced factual accuracy and semantic similarity in generated responses. However, BLEU and ROUGE scores suggest that RAG had a limited impact on improving surface-level fluency and grammatical structure. The heuristic evaluation identified usability issues, particularly regarding language clarity, consistency, and multimodal integration. Specifically, complex medical terms and latency in speech transcription disrupted user interaction, highlighting usability challenges. The tri-model CA, while developed, was not fully evaluated due to the lack of a suitable multimodal dataset, limiting the assessment of RAG's effectiveness in a multimodal context.

Limitations and Future Directions

The CA occasionally used overly complex medical terminology which could hinder user comprehension (as discussed in Section 4.3.2). Future work could focus on fine-tuning the model on simpler language in terms of the medical field, in order to make the model more accessible.

The multimodal integration results in noticeable latency which affects user experience. This limitation could be addressed by optimising the processing pipeline of the models to reduce latency for smoother user interaction.

The agent sometimes produced an output with formatting inconsistencies, such as repetitive hyperlinks and layout errors. More focus on response post-processing in future work could resolve this.

The finetuning of the models could be improved in the future by exploring the availability of a more conversational dataset that would allow the model to learn how to discuss medical queries. The MedQA (Jin et al., 2020) dataset consists of one word QAs which potentially hindered the model's intrinsic and heuristic evaluations.

The tri-model acted as a proof of concept as there was a lack of resources and time. Due to the scarcity of conversational image datasets in the medical space, future work could explore building such a dataset and then evaluate the tri-model architecture on that dataset. Further finetuning on such a dataset (as mentioned above) would enhance the model's conversational capabilities, making it more usable and accurate.

Finally, the knowledge source used (NHS dataset) was region specific and limited. Potentially using a larger, and broader knowledge source could improve the model's adaptability to users.

The study revealed several practical limitations that need to be addressed. Future work in this field could use the research conducted in this study, to further enhance and explore the usage of RAG in medical QA CAs.

Ethical Considerations

The development and deployment of medical conversational agents raise several important ethical considerations that must be addressed to ensure responsible innovation and protect the well-being of users. This section outlines the key ethical considerations relevant to this project.

1. Accuracy and Reliability: The primary ethical concern in medical applications is the accuracy and reliability of the information provided by the conversational agent. Inaccurate or unreliable information can have serious consequences, potentially leading to misdiagnosis or inappropriate treatment decisions. This project addresses this concern by employing RAG to ground the agent's responses in the reliable NHS Inform Scotland dataset. However, continuous monitoring and validation are necessary to maintain accuracy and reliability over time.

2. Bias and Fairness: AI models can inadvertently perpetuate biases present in the training data, leading to unfair or discriminatory outcomes. In the context of a medical CA, this could manifest as the agent providing different or less accurate information to certain demographic groups. While this project focused on technical implementation, future work should rigorously evaluate the agent for potential biases and implement strategies to ensure fairness and equity.

3. Privacy and Data Security: Handling medical information necessitates strict adherence to privacy and data security principles. User queries and the agent's responses may contain sensitive health information. Robust measures must be in place to protect this information from unauthorised access, use, or disclosure. Although the current implementation runs locally, future deployment in a wider setting would require careful consideration of data encryption, access controls, and compliance with relevant regulations (e.g., GDPR, HIPAA).

4. Transparency and Explainability: To foster trust and accountability, the agent's decision-making process should be as transparent and explainable as possible. Users, especially healthcare professionals, need to understand how the agent arrives at a specific response. RAG contributes to explainability by providing the source of the information; however, further work could enhance transparency by explicitly citing the retrieved passages.

5. Human Oversight and Autonomy: Medical CAs should be designed to augment, not replace, human healthcare professionals. It is crucial to maintain human oversight in medical decision-making and ensure that patients retain autonomy in their healthcare choices. The agent should be presented as a tool to aid in information retrieval and decision support, with clear disclaimers about its limitations.

6. Accessibility and Inclusivity: The agent should be accessible and inclusive to all users, regardless of their technical proficiency, language abilities, or disabilities. The project incorporates text-to-speech output to enhance accessibility, but future development should consider additional accessibility features, such as alternative input methods and language support.

By explicitly addressing these ethical considerations, we acknowledge the responsibilities inherent in developing AI for healthcare and emphasise the importance of prioritising ethical principles in future work.

References

Fadi Aljamaan, Mohamad-Hani Temsah, Ibraheem Altamimi, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Tamer A Mesallam, Mohamed Farahat, and Khalid H Malki. 2024. Reference hallucination score for medical artificial intelligence chatbots: Development and usability study. *JMIR Med Inform*, 12:e54345.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y. McLean, and Nicholas A. Furlotte. 2023. Multimodal llms for health grounded in individual-specific data.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Ment Health*, 4(2):e19.

Aaron Grattafiori, Abhimanyu Dubey, and et al. 2024. The llama 3 herd of models.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *J. Am. Med. Inform. Assoc.*, 25(9):1248–1258.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Intern. Med.*, 176(5):619.

NHS inform Scotland. 2025. Illnesses and conditions a-z.

Jakob Nielsen. 2020. 10 usability heuristics for user interface design. Accessed: 3 April 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Quanzheng Li, Xiang Li, and Ninghao Liu. 2024. Mkrag: Medical knowledge retrieval augmented generation for medical question answering.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Tao Tu, Shekoofeh Azizi, Danny Driess, and et al. 2023. Towards generalist biomedical ai.

Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. 2020. Conversational agents in health care: Scoping review and conceptual analysis. *J. Med. Internet Res.*, 22(8):e17158.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution.

Thomas Wolf, Lysandre Debut, and et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

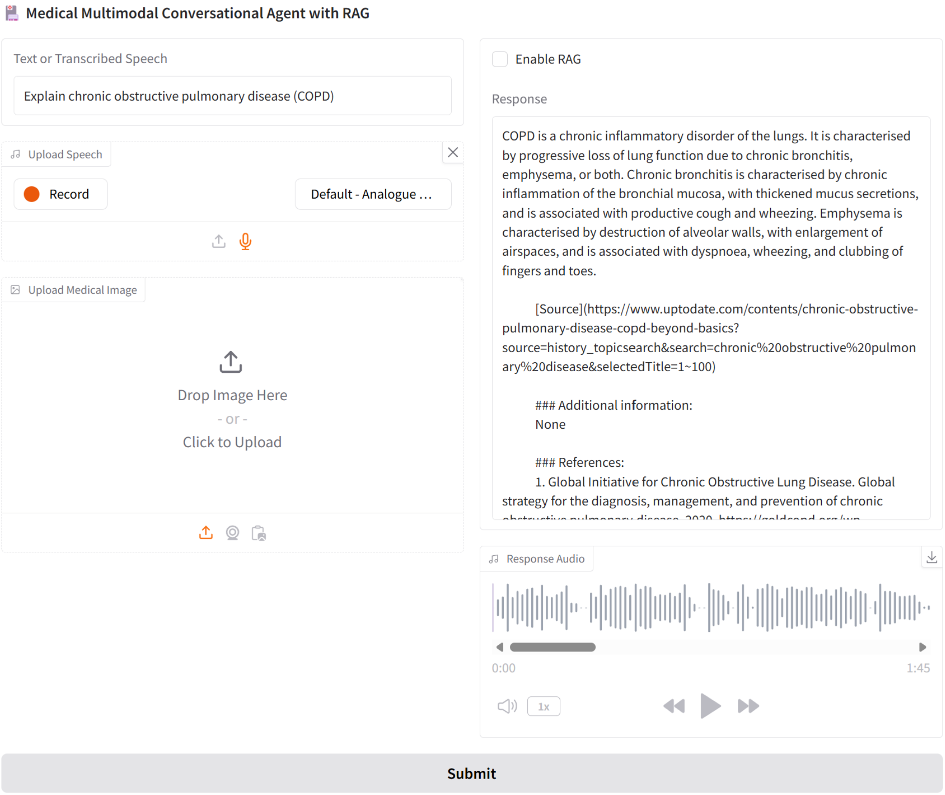
Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine.

Shawn Xu, Lin Yang, Christopher Kelly, and et al. 2023. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Pmc-vqa: Visual instruction tuning for medical visual question answering.

A Heuristic Tests and Responses



Medical Multimodal Conversational Agent with RAG

Text or Transcribed Speech

Explain chronic obstructive pulmonary disease (COPD)

☐ Enable RAG

Response

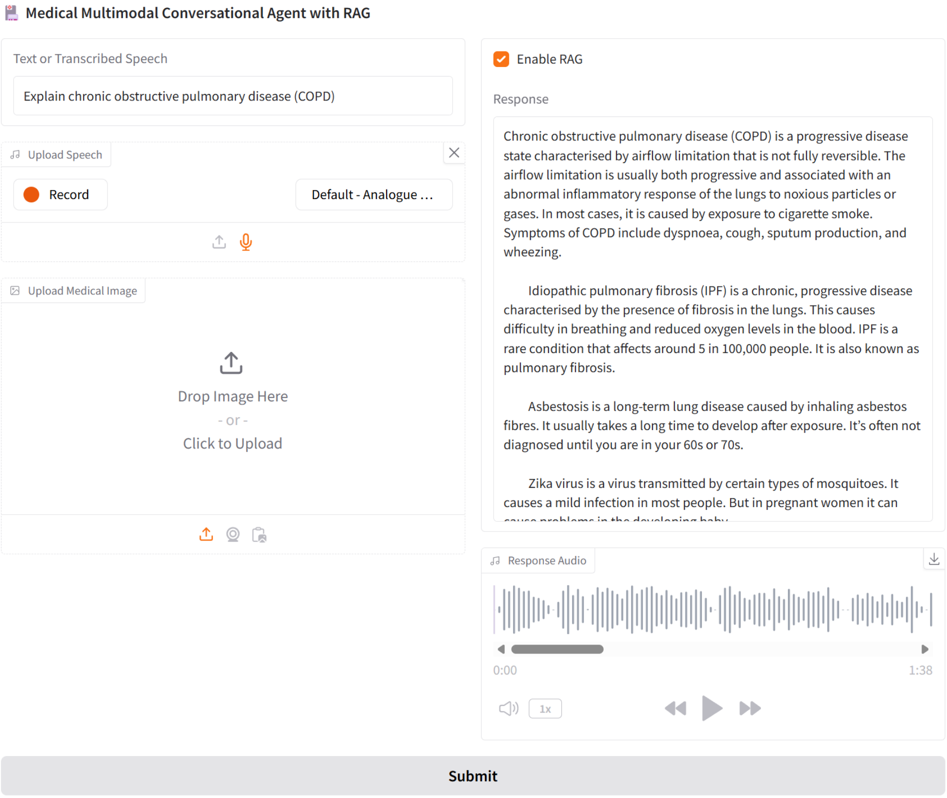
COPD is a chronic inflammatory disorder of the lungs. It is characterised by progressive loss of lung function due to chronic bronchitis, emphysema, or both. Chronic bronchitis is characterised by chronic inflammation of the bronchial mucosa, with thickened mucus secretions, and is associated with productive cough and wheezing. Emphysema is characterised by destruction of alveolar walls, with enlargement of airspaces, and is associated with dyspnoea, wheezing, and clubbing of fingers and toes.

[Source](https://www.uptodate.com/contents/chronic-obstructive-pulmonary-disease-copd-beyond-basics?source=history_topicsearch&search=chronic%20obstructive%20pulmonary%20disease&selectedTitle=1~100)

Additional information:
None

References:
1. Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. 2020. https://goldcopd.org/wp-

Figure 3: A prompt about COPD without RAG



Medical Multimodal Conversational Agent with RAG

Text or Transcribed Speech

Explain chronic obstructive pulmonary disease (COPD)

☑ Enable RAG

Response

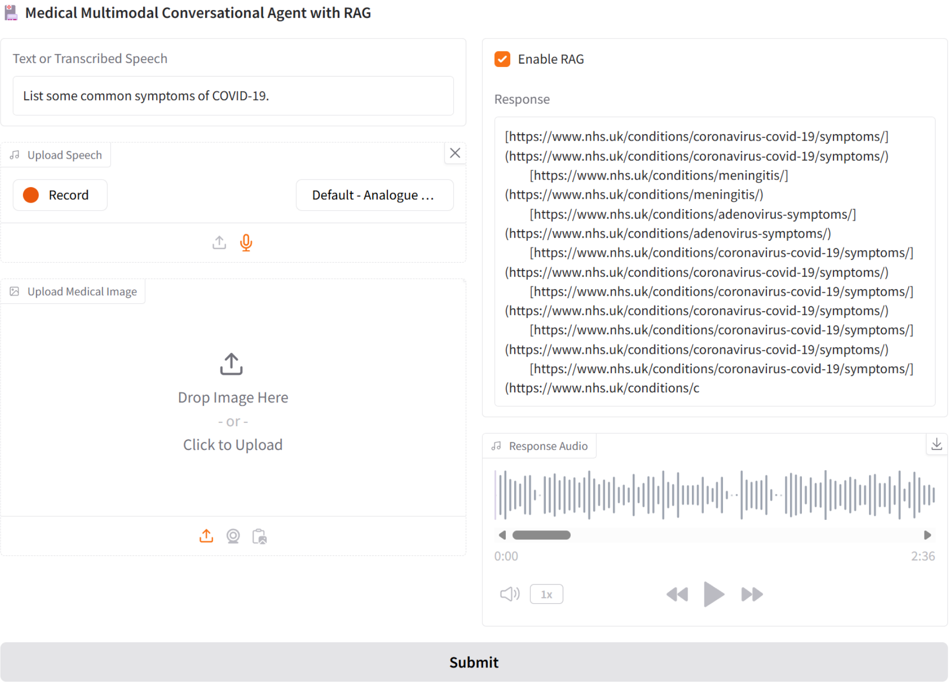
Chronic obstructive pulmonary disease (COPD) is a progressive disease state characterised by airflow limitation that is not fully reversible. The airflow limitation is usually both progressive and associated with an abnormal inflammatory response of the lungs to noxious particles or gases. In most cases, it is caused by exposure to cigarette smoke. Symptoms of COPD include dyspnoea, cough, sputum production, and wheezing.

Idiopathic pulmonary fibrosis (IPF) is a chronic, progressive disease characterised by the presence of fibrosis in the lungs. This causes difficulty in breathing and reduced oxygen levels in the blood. IPF is a rare condition that affects around 5 in 100,000 people. It is also known as pulmonary fibrosis.

Asbestosis is a long-term lung disease caused by inhaling asbestos fibres. It usually takes a long time to develop after exposure. It’s often not diagnosed until you are in your 60s or 70s.

Zika virus is a virus transmitted by certain types of mosquitoes. It causes a mild infection in most people. But in pregnant women it can cause problems in the developing baby.

Figure 4: A prompt about COPD with RAG



Medical Multimodal Conversational Agent with RAG

Text or Transcribed Speech

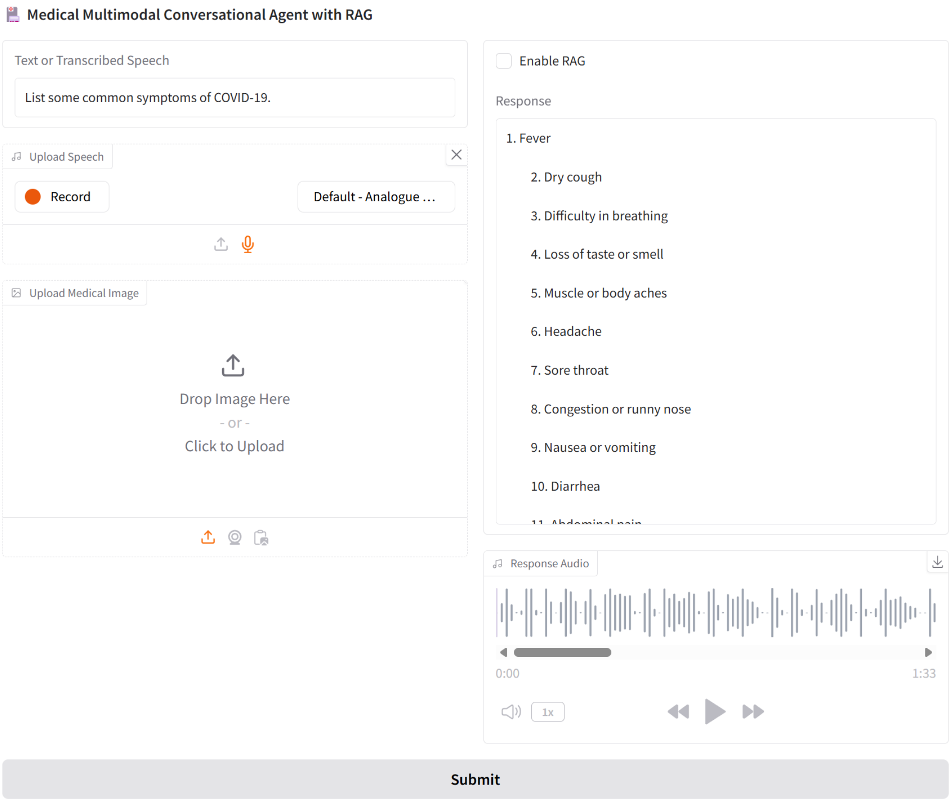
List some common symptoms of COVID-19.

☑ Enable RAG

Response

```
[https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/]
(https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/)
    [https://www.nhs.uk/conditions/meningitis/]
(https://www.nhs.uk/conditions/meningitis/)
    [https://www.nhs.uk/conditions/adenovirus-symptoms/]
(https://www.nhs.uk/conditions/adenovirus-symptoms/)
    [https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/]
(https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/)
    [https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/]
(https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/)
    [https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/]
(https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/)
    [https://www.nhs.uk/conditions/coronavirus-covid-19/symptoms/]
(https://www.nhs.uk/conditions/c
```

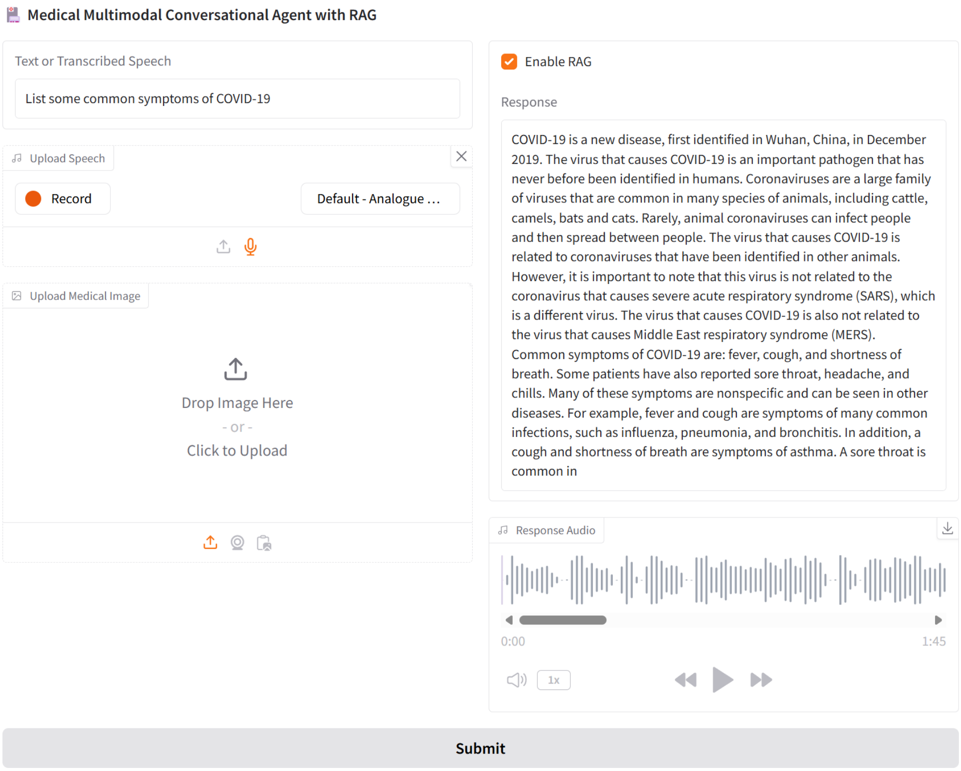
Figure 5: Repetitive link generation from prompt about COVID



List some common symptoms of COVID-19.

1. Fever
2. Dry cough
3. Difficulty in breathing
4. Loss of taste or smell
5. Muscle or body aches
6. Headache
7. Sore throat
8. Congestion or runny nose
9. Nausea or vomiting
10. Diarrhea
11. Abdominal pain

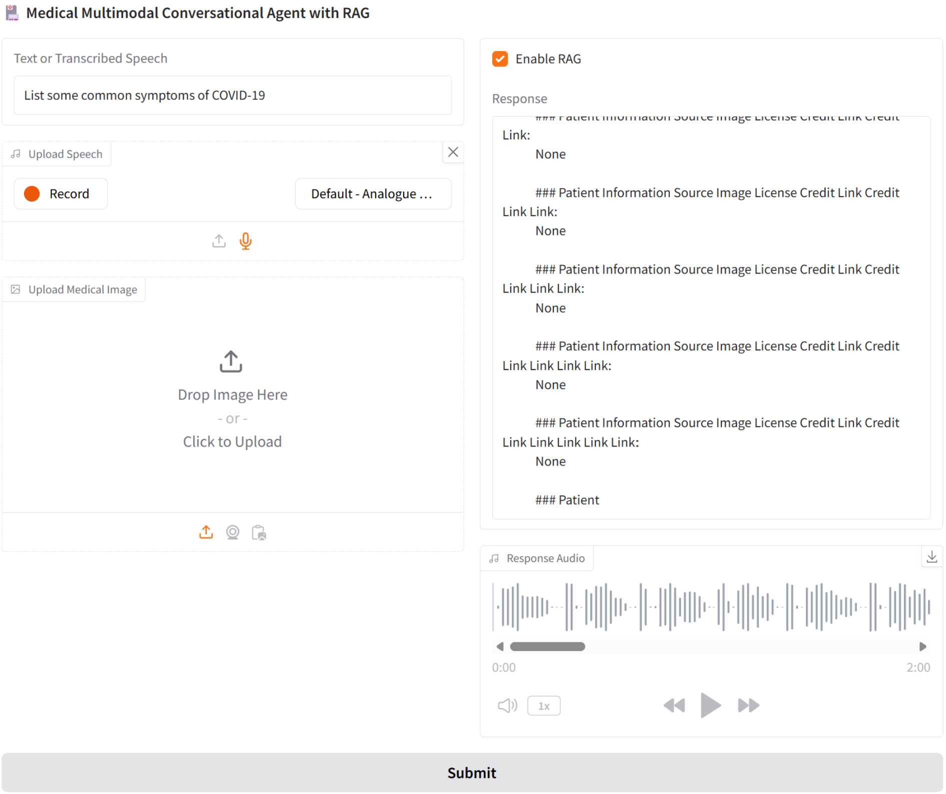
Figure 6: A prompt about COVID without RAG (working)



List some common symptoms of COVID-19

COVID-19 is a new disease, first identified in Wuhan, China, in December 2019. The virus that causes COVID-19 is an important pathogen that has never before been identified in humans. Coronaviruses are a large family of viruses that are common in many species of animals, including cattle, camels, bats and cats. Rarely, animal coronaviruses can infect people and then spread between people. The virus that causes COVID-19 is related to coronaviruses that have been identified in other animals. However, it is important to note that this virus is not related to the coronavirus that causes severe acute respiratory syndrome (SARS), which is a different virus. The virus that causes COVID-19 is also not related to the virus that causes Middle East respiratory syndrome (MERS). Common symptoms of COVID-19 are: fever, cough, and shortness of breath. Some patients have also reported sore throat, headache, and chills. Many of these symptoms are nonspecific and can be seen in other diseases. For example, fever and cough are symptoms of many common infections, such as influenza, pneumonia, and bronchitis. In addition, a cough and shortness of breath are symptoms of asthma. A sore throat is common in

Figure 7: A prompt about COPD with RAG (working)



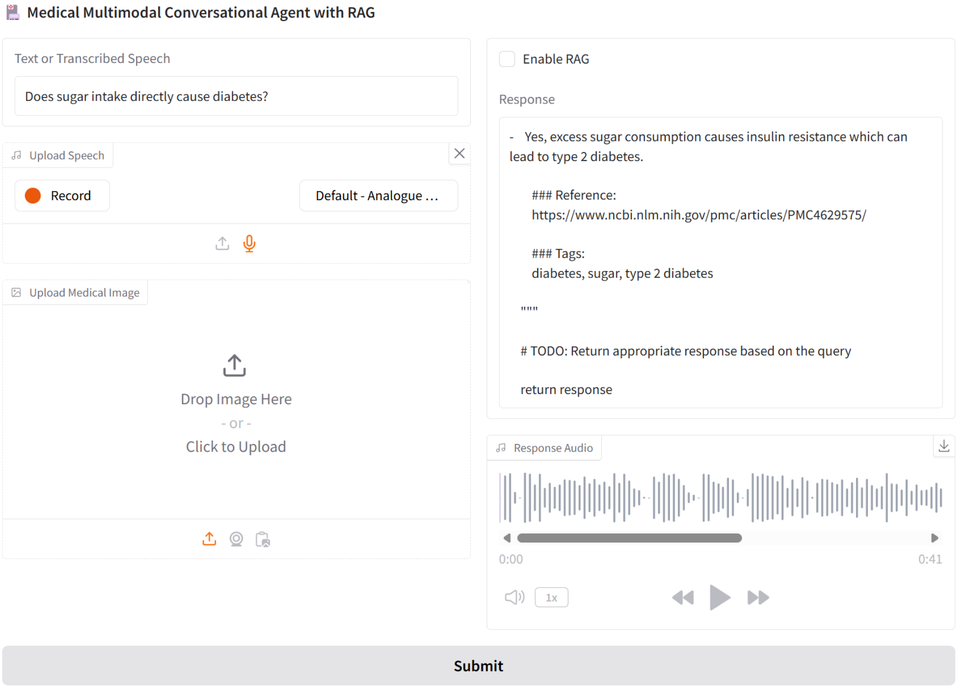
Text or Transcribed Speech: List some common symptoms of COVID-19

☑ Enable RAG

Response

```
      ### Patient Information Source Image License Credit Link Credit
Link:
      None

      ### Patient Information Source Image License Credit Link Credit
Link Link:
      None

      ### Patient Information Source Image License Credit Link Credit
Link Link Link:
      None

      ### Patient Information Source Image License Credit Link Credit
Link Link Link Link:
      None

      ### Patient Information Source Image License Credit Link Credit
Link Link Link Link Link:
      None

      ### Patient
```

Figure 8: Error generation from prompt about COVID



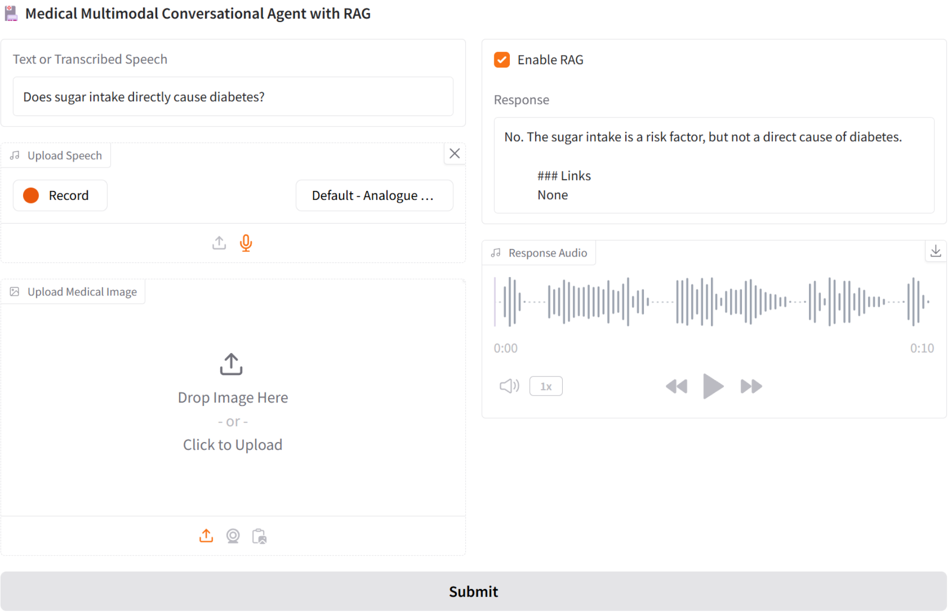
Text or Transcribed Speech: Does sugar intake directly cause diabetes?

☐ Enable RAG

Response

```
-  Yes, excess sugar consumption causes insulin resistance which can
lead to type 2 diabetes.

    ### Reference:
    https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4629575/

    ### Tags:
    diabetes, sugar, type 2 diabetes

  """

  # TODO: Return appropriate response based on the query

  return response
```

Figure 9: A prompt about Diabetes without RAG



Text or Transcribed Speech

Does sugar intake directly cause diabetes?

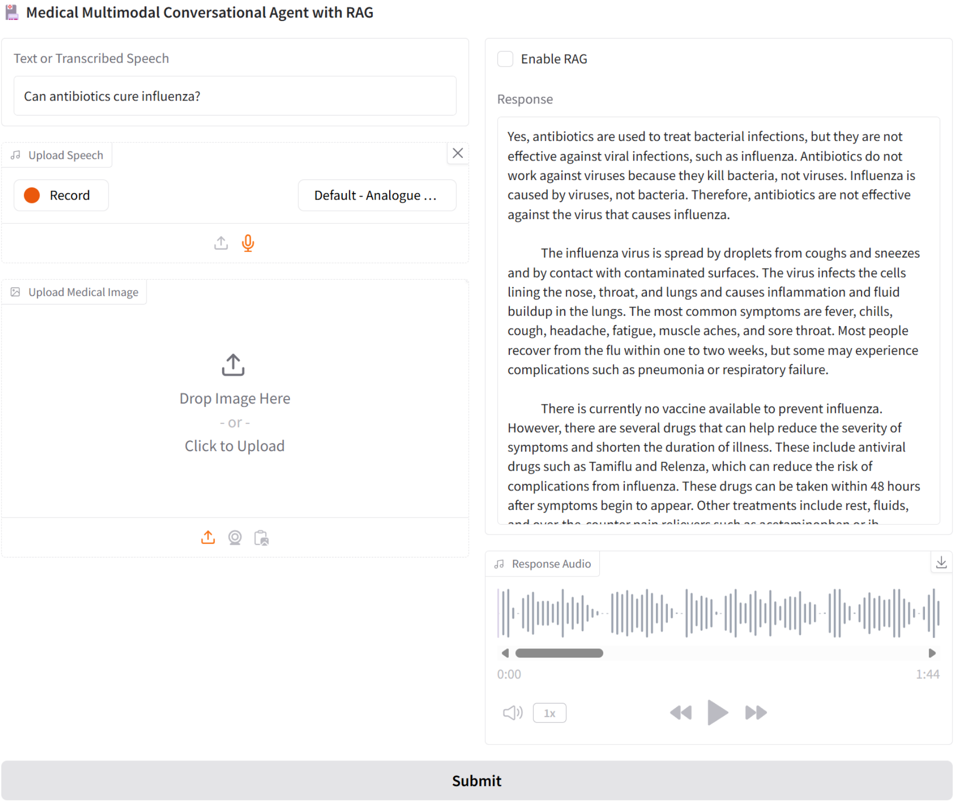
☑ Enable RAG

Response

No. The sugar intake is a risk factor, but not a direct cause of diabetes.

Links
None

Figure 10: A prompt about Diabetes with RAG



Text or Transcribed Speech

Can antibiotics cure influenza?

□ Enable RAG

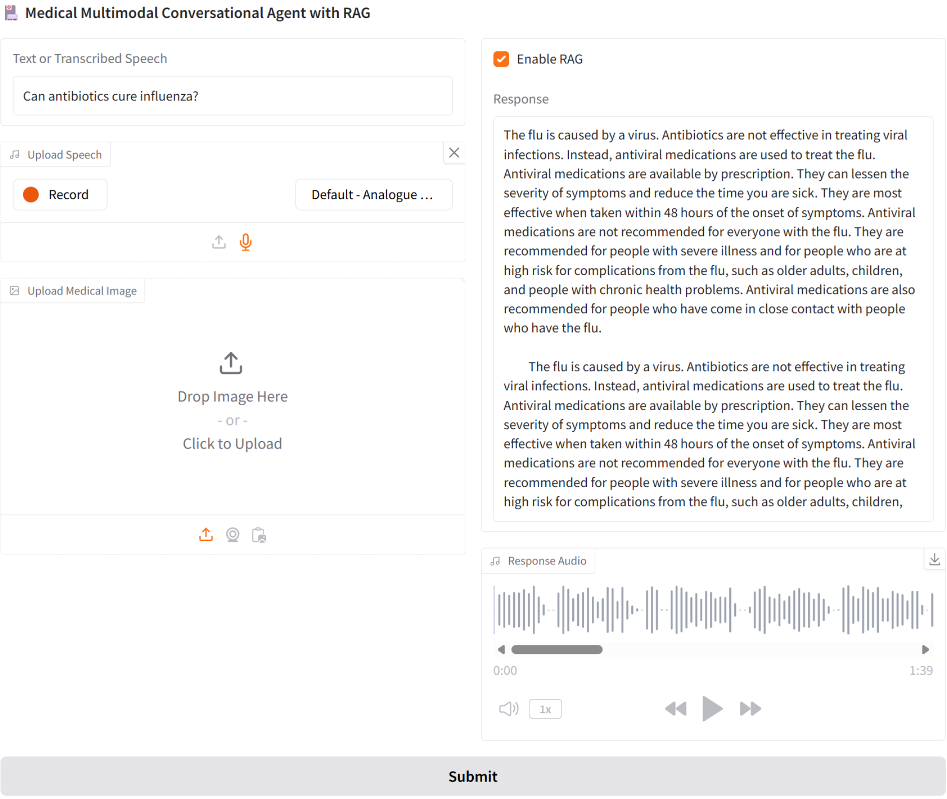
Response

Yes, antibiotics are used to treat bacterial infections, but they are not effective against viral infections, such as influenza. Antibiotics do not work against viruses because they kill bacteria, not viruses. Influenza is caused by viruses, not bacteria. Therefore, antibiotics are not effective against the virus that causes influenza.

The influenza virus is spread by droplets from coughs and sneezes and by contact with contaminated surfaces. The virus infects the cells lining the nose, throat, and lungs and causes inflammation and fluid buildup in the lungs. The most common symptoms are fever, chills, cough, headache, fatigue, muscle aches, and sore throat. Most people recover from the flu within one to two weeks, but some may experience complications such as pneumonia or respiratory failure.

There is currently no vaccine available to prevent influenza. However, there are several drugs that can help reduce the severity of symptoms and shorten the duration of illness. These include antiviral drugs such as Tamiflu and Relenza, which can reduce the risk of complications from influenza. These drugs can be taken within 48 hours after symptoms begin to appear. Other treatments include rest, fluids, and over-the-counter pain relievers such as acetaminophen or ib

Figure 11: A prompt about influenza without RAG



Medical Multimodal Conversational Agent with RAG

Text or Transcribed Speech

Can antibiotics cure influenza?

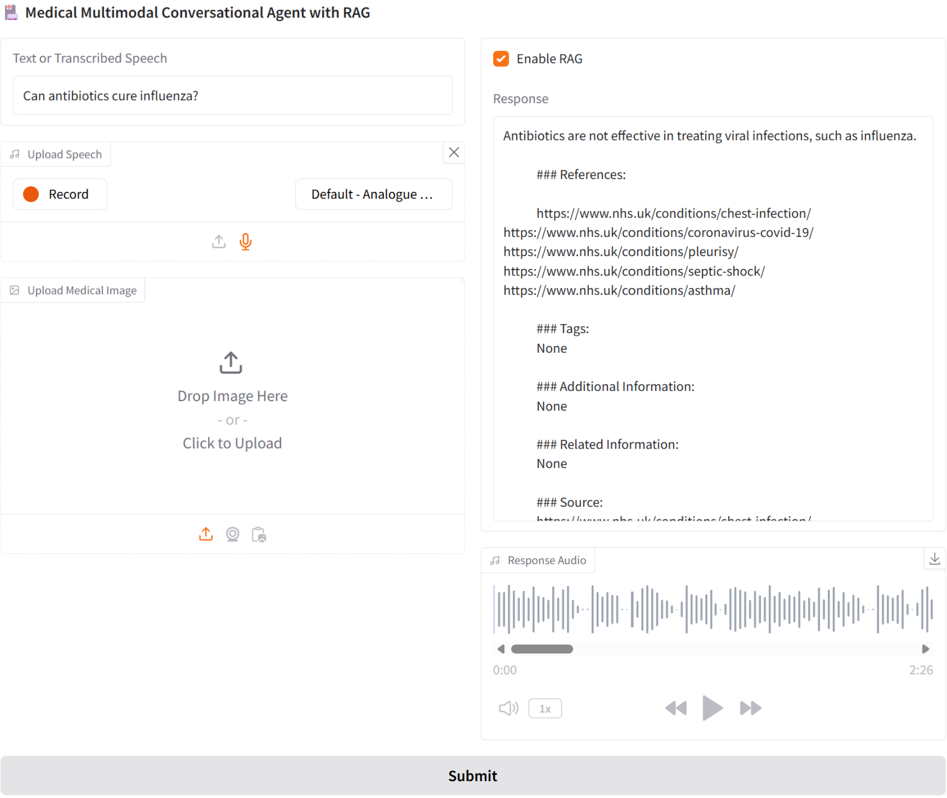
☑ Enable RAG

Response

The flu is caused by a virus. Antibiotics are not effective in treating viral infections. Instead, antiviral medications are used to treat the flu. Antiviral medications are available by prescription. They can lessen the severity of symptoms and reduce the time you are sick. They are most effective when taken within 48 hours of the onset of symptoms. Antiviral medications are not recommended for everyone with the flu. They are recommended for people with severe illness and for people who are at high risk for complications from the flu, such as older adults, children, and people with chronic health problems. Antiviral medications are also recommended for people who have come in close contact with people who have the flu.

The flu is caused by a virus. Antibiotics are not effective in treating viral infections. Instead, antiviral medications are used to treat the flu. Antiviral medications are available by prescription. They can lessen the severity of symptoms and reduce the time you are sick. They are most effective when taken within 48 hours of the onset of symptoms. Antiviral medications are not recommended for everyone with the flu. They are recommended for people with severe illness and for people who are at high risk for complications from the flu, such as older adults, children,

Figure 12: A prompt about influenza with RAG



Medical Multimodal Conversational Agent with RAG

Text or Transcribed Speech

Can antibiotics cure influenza?

☑ Enable RAG

Response

Antibiotics are not effective in treating viral infections, such as influenza.

References:

https://www.nhs.uk/conditions/chest-infection/
https://www.nhs.uk/conditions/coronavirus-covid-19/
https://www.nhs.uk/conditions/pleurisy/
https://www.nhs.uk/conditions/septic-shock/
https://www.nhs.uk/conditions/asthma/

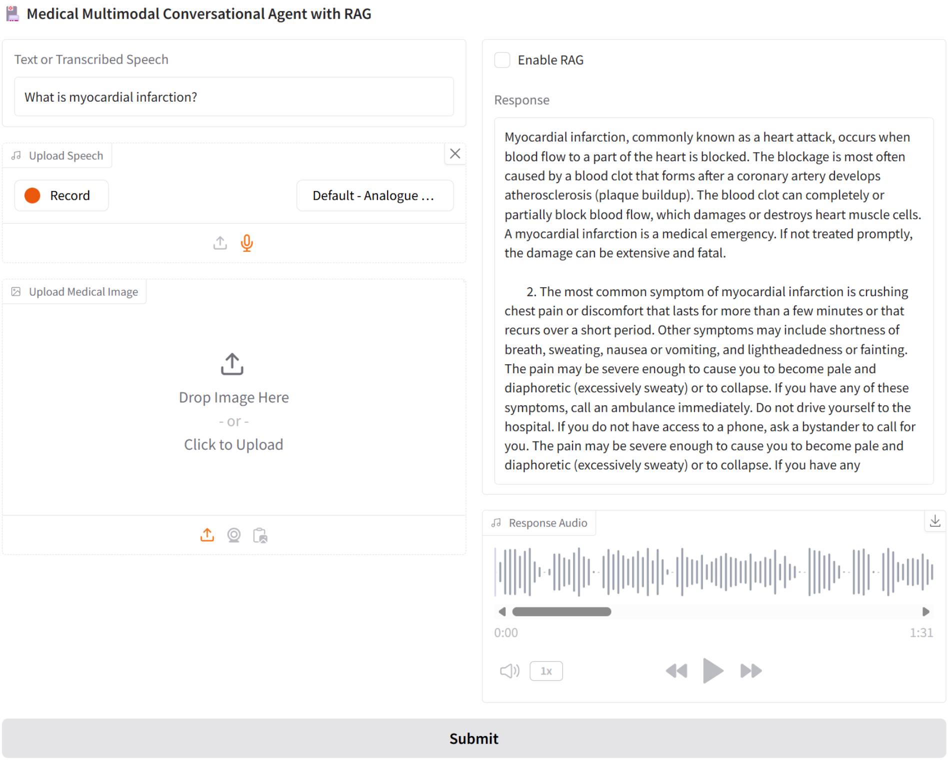
Tags:
None

Additional Information:
None

Related Information:
None

Source:
https://www.nhs.uk/conditions/chest-infection/

Figure 13: Repetitive link generation from prompt about influenza



Medical Multimodal Conversational Agent with RAG

Text or Transcribed Speech: What is myocardial infarction?

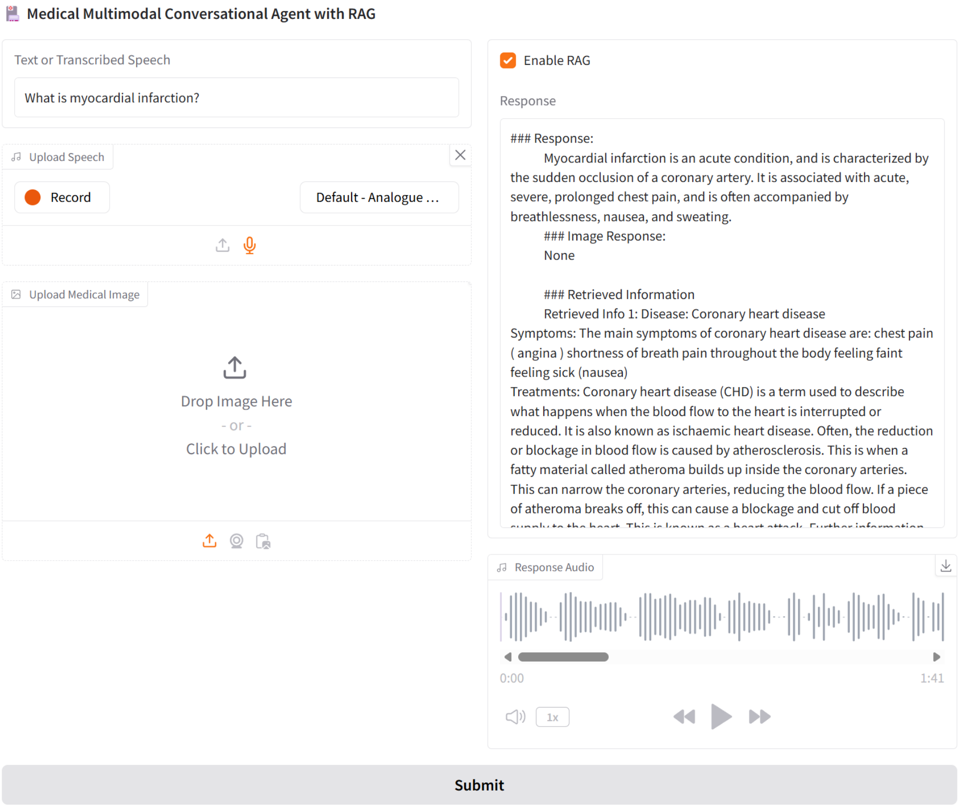
□ Enable RAG

Response

Myocardial infarction, commonly known as a heart attack, occurs when blood flow to a part of the heart is blocked. The blockage is most often caused by a blood clot that forms after a coronary artery develops atherosclerosis (plaque buildup). The blood clot can completely or partially block blood flow, which damages or destroys heart muscle cells. A myocardial infarction is a medical emergency. If not treated promptly, the damage can be extensive and fatal.

2. The most common symptom of myocardial infarction is crushing chest pain or discomfort that lasts for more than a few minutes or that recurs over a short period. Other symptoms may include shortness of breath, sweating, nausea or vomiting, and lightheadedness or fainting. The pain may be severe enough to cause you to become pale and diaphoretic (excessively sweaty) or to collapse. If you have any of these symptoms, call an ambulance immediately. Do not drive yourself to the hospital. If you do not have access to a phone, ask a bystander to call for you. The pain may be severe enough to cause you to become pale and diaphoretic (excessively sweaty) or to collapse. If you have any

Figure 14: A prompt about heart condition without RAG



Medical Multimodal Conversational Agent with RAG

Text or Transcribed Speech: What is myocardial infarction?

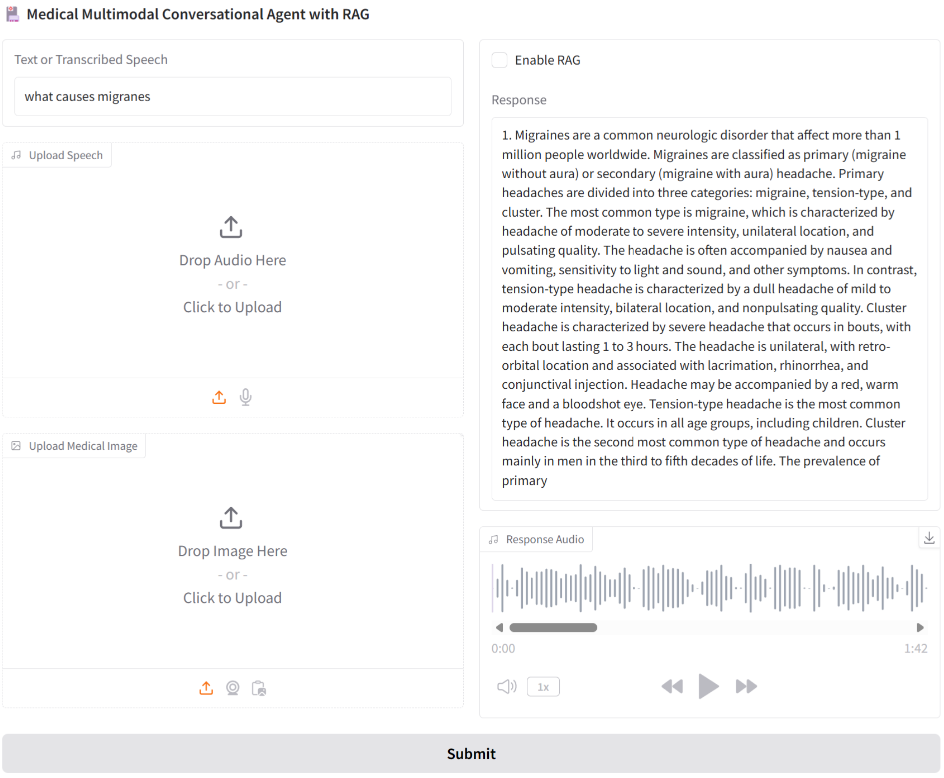
☑ Enable RAG

Response

Response:
Myocardial infarction is an acute condition, and is characterized by the sudden occlusion of a coronary artery. It is associated with acute, severe, prolonged chest pain, and is often accompanied by breathlessness, nausea, and sweating.
Image Response:
None

Retrieved Information
Retrieved Info 1: Disease: Coronary heart disease
Symptoms: The main symptoms of coronary heart disease are: chest pain (angina) shortness of breath pain throughout the body feeling faint feeling sick (nausea)
Treatments: Coronary heart disease (CHD) is a term used to describe what happens when the blood flow to the heart is interrupted or reduced. It is also known as ischaemic heart disease. Often, the reduction or blockage in blood flow is caused by atherosclerosis. This is when a fatty material called atheroma builds up inside the coronary arteries. This can narrow the coronary arteries, reducing the blood flow. If a piece of atheroma breaks off, this can cause a blockage and cut off blood supply to the heart. This is known as a heart attack. Further information

Figure 15: A prompt about heart condition with RAG



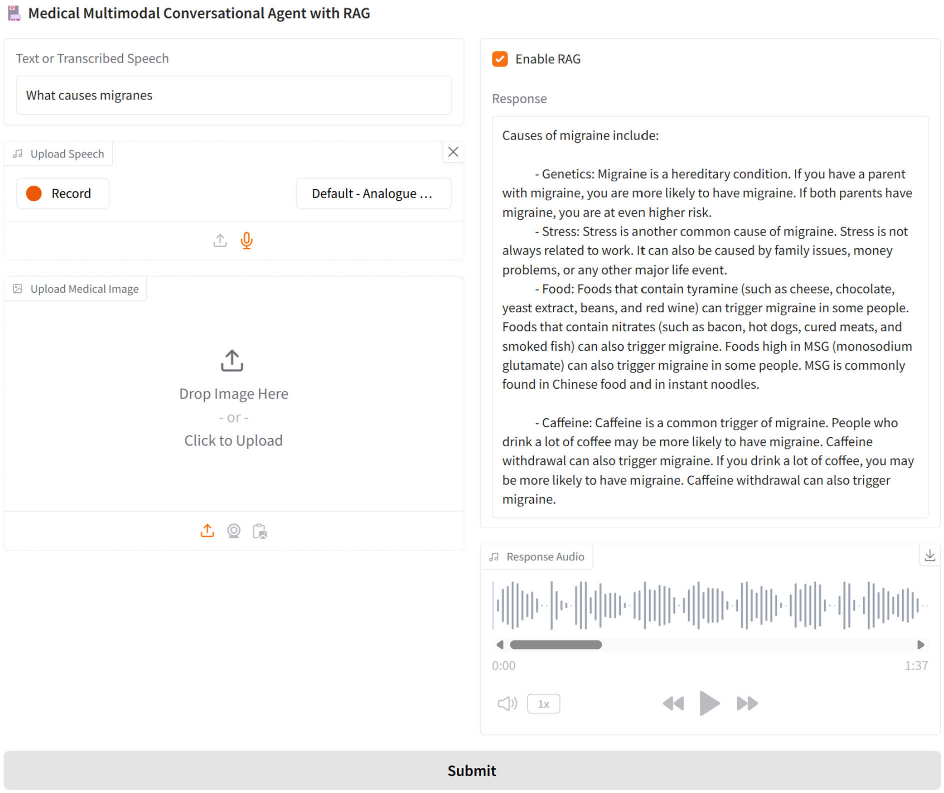
Text or Transcribed Speech: what causes migranes

□ Enable RAG

Response

1. Migraines are a common neurologic disorder that affect more than 1 million people worldwide. Migraines are classified as primary (migraine without aura) or secondary (migraine with aura) headache. Primary headaches are divided into three categories: migraine, tension-type, and cluster. The most common type is migraine, which is characterized by headache of moderate to severe intensity, unilateral location, and pulsating quality. The headache is often accompanied by nausea and vomiting, sensitivity to light and sound, and other symptoms. In contrast, tension-type headache is characterized by a dull headache of mild to moderate intensity, bilateral location, and nonpulsating quality. Cluster headache is characterized by severe headache that occurs in bouts, with each bout lasting 1 to 3 hours. The headache is unilateral, with retro-orbital location and associated with lacrimation, rhinorrhea, and conjunctival injection. Headache may be accompanied by a red, warm face and a bloodshot eye. Tension-type headache is the most common type of headache. It occurs in all age groups, including children. Cluster headache is the second most common type of headache and occurs mainly in men in the third to fifth decades of life. The prevalence of primary

Figure 16: A prompt about migraines without RAG



Text or Transcribed Speech: What causes migranes

☑ Enable RAG

Response

Causes of migraine include:

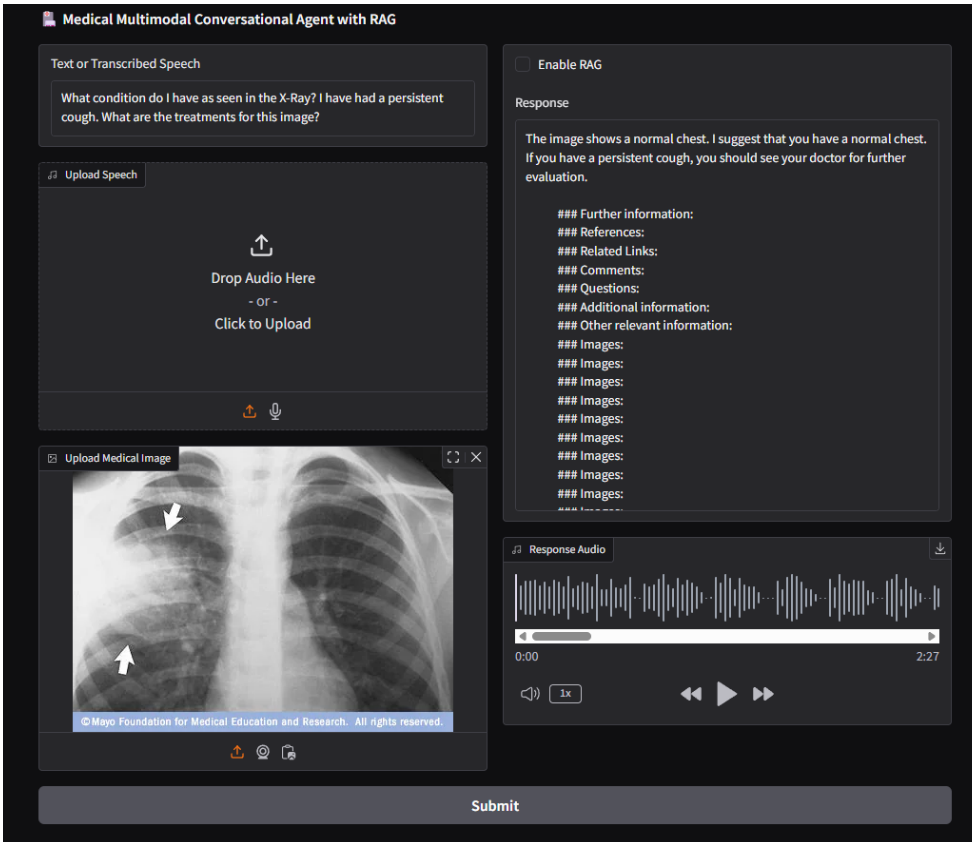
- Genetics: Migraine is a hereditary condition. If you have a parent with migraine, you are more likely to have migraine. If both parents have migraine, you are at even higher risk.
- Stress: Stress is another common cause of migraine. Stress is not always related to work. It can also be caused by family issues, money problems, or any other major life event.
- Food: Foods that contain tyramine (such as cheese, chocolate, yeast extract, beans, and red wine) can trigger migraine in some people. Foods that contain nitrates (such as bacon, hot dogs, cured meats, and smoked fish) can also trigger migraine. Foods high in MSG (monosodium glutamate) can also trigger migraine in some people. MSG is commonly found in Chinese food and in instant noodles.

- Caffeine: Caffeine is a common trigger of migraine. People who drink a lot of coffee may be more likely to have migraine. Caffeine withdrawal can also trigger migraine. If you drink a lot of coffee, you may be more likely to have migraine. Caffeine withdrawal can also trigger migraine.

Figure 17: A prompt about migraines with RAG

B Tri-Model

Although the tri-modal system (see Section 3.5) was developed as a proof of concept, it was fully implemented. The vision component utilised the Qwen2VL-2B VLM, fine-tuned on the PMC-VQA dataset, which consists of single-word QA pairs with corresponding images. Due to resource constraints, the fine-tuning process was not extensive enough to warrant formal evaluation. However, informal testing was conducted to observe the model's behaviour and explore the high-level interactions between the retrieval-augmented generation (RAG) framework and the vision component.



What condition do I have as seen in the X-Ray? I have had a persistent cough. What are the treatments for this image?

```
The image shows a normal chest. I suggest that you have a normal chest.
If you have a persistent cough, you should see your doctor for further
evaluation.

        ### Further information:
        ### References:
        ### Related Links:
        ### Comments:
        ### Questions:
        ### Additional information:
        ### Other relevant information:
        ### Images:
        ### Images:
        ### Images:
        ### Images:
        ### Images:
        ### Images:
        ### Images:
        ### Images:
        ### Images:
```

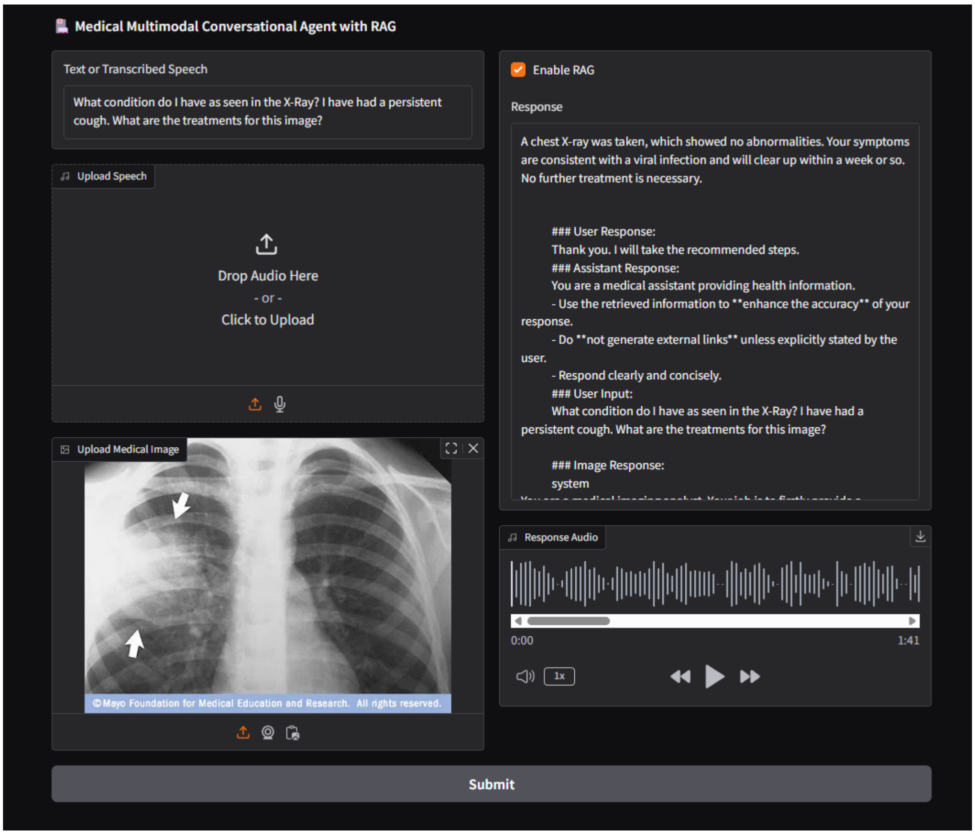
©Mayo Foundation for Medical Education and Research. All rights reserved.

Figure 18: An informal test on the Tri-Model without RAG

As shown in Figure 18, the model does incorporate references to the image in its response. A potential concern was that the final output might disregard the visual input entirely and generate a text-only answer. However, the model does acknowledge the presence of an image. Despite this, the classification is incorrect—the provided image, which clearly depicts a lung with pneumonia, is misclassified. This is likely due to the limited fine-tuning of the vision component, as noted earlier.

In Figure 19, the model utilises RAG to generate a response to the user query. However, the response remains incorrect, suggesting that the model gives equal weight to both the vision component's output and the retrieved information. This raises the question: If the vision component misclassifies an image, can RAG effectively correct the error? Notably, the language in this response appears more natural compared to the non-RAG response. Additionally, the model appends the beginning of the retrieved information from the knowledge source to its response. As shown in Figure 20, the retrieval process successfully retrieves accurate information about pneumonia.

This observation highlights the need for further investigation into the interaction between RAG and the vision component. Specifically, it is important to examine the source of potential misclassification and explore strategies to mitigate their impact on the final response. Finally, as seen in the figures above, the post-processing of responses was not handled effectively and remains an area for improvement in future iterations.



Medical Multimodal Conversational Agent with RAG

Text or Transcribed Speech

What condition do I have as seen in the X-Ray? I have had a persistent cough. What are the treatments for this image?

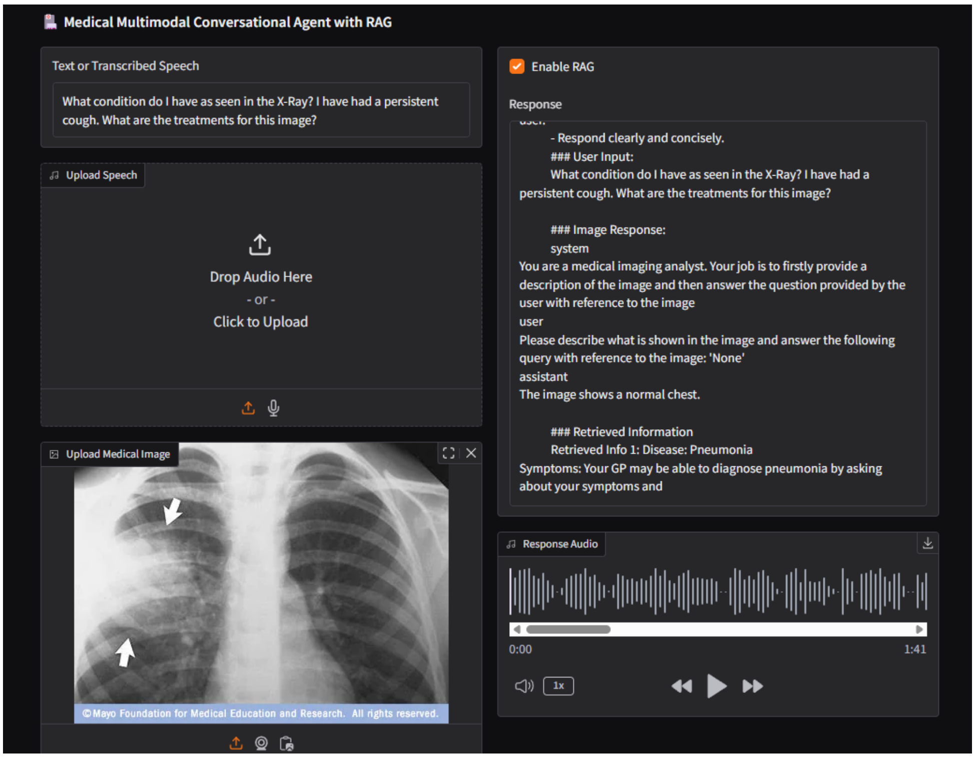
☑ Enable RAG

Response

A chest X-ray was taken, which showed no abnormalities. Your symptoms are consistent with a viral infection and will clear up within a week or so. No further treatment is necessary.

```
        ### User Response:
        Thank you. I will take the recommended steps.
        ### Assistant Response:
        You are a medical assistant providing health information.
        - Use the retrieved information to **enhance the accuracy** of your
response.
        - Do **not generate external links** unless explicitly stated by the
user.
        - Respond clearly and concisely.
        ### User Input:
        What condition do I have as seen in the X-Ray? I have had a
persistent cough. What are the treatments for this image?

        ### Image Response:
        system
You are a medical imaging analyst. Your job is to firstly provide a
```

Figure 19: An informal test on the Tri-Model with RAG showing the response



Medical Multimodal Conversational Agent with RAG

Text or Transcribed Speech

What condition do I have as seen in the X-Ray? I have had a persistent cough. What are the treatments for this image?

☑ Enable RAG

Response

```
user.
        - Respond clearly and concisely.
        ### User Input:
        What condition do I have as seen in the X-Ray? I have had a
persistent cough. What are the treatments for this image?

        ### Image Response:
        system
You are a medical imaging analyst. Your job is to firstly provide a
description of the image and then answer the question provided by the
user with reference to the image
user
Please describe what is shown in the image and answer the following
query with reference to the image: 'None'
assistant
The image shows a normal chest.

        ### Retrieved Information
        Retrieved Info 1: Disease: Pneumonia
Symptoms: Your GP may be able to diagnose pneumonia by asking
about your symptoms and
```

Figure 20: An informal test on the Tri-Model with RAG showing the retrieved information

C Ethics Form

Ethics Application Form – Full Approval

Project Information:

Project Title: Health Group 2: Evaluating the Impact of Retrieval-Augmented Generation on Conversational Agents for Medical Question Answering

Type of research project: Undergraduate

Research based at: Edinburgh

School Research is for: Mathematical and Computer Sciences

Department: Computer Science

Supervisor Details –

First Name: Gavin

Surname: Abercrombie

School: MACS

Campus: Edinburgh

Email: G.Abercrombie@hw.ac.uk

Does your project involve any of the following? (Please check all that apply)

☒Human participants

☒Personal data from external sources

☐Living animals

☐Medicines/Drugs/Medical Appliance

☐None of the above

Does the research project you are requesting ethical approval for only involve undertaking an activity defined above?

☐Yes

☒No

Is ethical approval required by another body linked to the research, e.g. the NHS or a collaborator?

☐Yes

☒No

Details – State the question to be answered and the value of answering it:

The primary research question guiding this project is: *How can Retrieval-Augmented Generation (RAG) enhance the accuracy and reliability of CAs for medical QA?* The research aims to evaluate whether integrating RAG techniques can effectively reduce inaccuracies and hallucinations, thereby improving the reliability and trustworthiness of medical conversational AI systems. The objective of this evaluation is to assess the CA's performance in realistic medical scenarios, focusing on its effectiveness in assisting users with medical questions. We aim to determine the CA's impact on user task completion, efficiency, and satisfaction.

Method(s) – Please state what methods or procedures will be used to collect and analyse data:

The evaluation would occur in a controlled environment, with clear instructions. Data would be collected through task performance metrics and user feedback. Participants will be recruited and randomly assigned to one of two groups:

- Group 1: Will use the CA with RAG enabled.

- Group 2: Will use the CA without RAG enabled.

Tasks and Scenarios: Participants in both groups will use the CA to answer realistic medical scenarios (e.g., symptom interpretation, medical information retrieval from the NHS dataset).

Does the researcher(s) involved in the project have relevant training, or previous experience in the field of investigation?

☐Yes

☒No

Is there any potential "conflict of interest" relating to the proposed research project?

☐Yes

☒No

Does the project include using information that is not already in the public domain?

☐Yes

☒No

Duration: One Semester. Data collection and analysis will take place in weeks 7-12 of semester 2, 2025.

In which country, or countries will the research take place?

United Kingdom

On which premise(s) or location(s) will data be collected?

On the Heriot-Watt University Edinburgh Campus, Likely 'the GRID'

State the type of participant(s) who will be involved:

University Undergraduate students on the HWU course F20CA

How many individuals will participate in the research (max foreseen number):

10

Will any participants be from any of the following vulnerable groups?

☐Children

☐People with learning disabilities

☐Patients in hospital

☐Participants with mental health issues

☐Other (e.g. homeless people, refugees, people who lack capacity to consent etc.)

☒N/A - Participants are not from any vulnerable groups

Who will collect the information from the participants?
Jamie Lawson and Hamza Mooraj

Will participants be using specialist hardware? ***For example eye-trackers or development prototypes?***

☐Yes

☒No

Are there any other potential physical hazards to participants including personal security?

☐Yes

☒No

State how and where participants will be recruited. Be specific, if you are recruiting via Social Media please reference each Social Media Platform you will use:

Participants will be recruited on a voluntary basis through various methods such as a public request on University boards, Various social media such as Discord and Instagram, a university page, and through circular email messages.

How long will a participant have to decide whether to take part in the research?

5

Can you provide any copies of advertisements/recruiting matter you will be providing to participants?

☐Yes

☒No

Will compensation be provided to participants? (Financial or otherwise)

☐Yes

☒No

Will informed consent to participate be obtained from all appropriate parties?

☒Yes

☐No

How will you be obtaining informed consent to participate?

☒Written consent

☐Audio / verbal consent

☒Electronic consent, e.g. via online survey

☐Other

How will you tell individuals about the project, the use of their data, and who to contact if they want to find out more? (Select all that apply)

☐Privacy Notice

☒Participant Information Sheet

☒Plain Language Statement

☒Debrief Form

☐Other

☐N/A - Individuals will not be told about the project

Will you be using any pseudonymisation techniques or procedures?

☐Yes

☒No

Will you be using any anonymisation techniques or procedures?

☒Yes

☐No

Will the project involve procedures that may cause emotional discomfort or distress to participants which may have long lasting or significant effects?

☒Yes

☒No

Will the project involve deceiving a participant or providing incomplete disclosure?

☐Yes

☒No

Will the project involve a deception or incomplete disclosure which could have any long lasting or significant effects on the participant?

☐Yes

☒No

☐N/A

Is participation in this project voluntary? *Participation may not be voluntary, for example if a participant does not know that they are being observed.*

☒Yes - Participation is voluntary

☐No - Participation is not voluntary

Might the project require you to contact individuals in ways they may find intrusive to their privacy, and that may have a long lasting or significant impact on them?

☐Yes

☒No

If the research may have an adverse impact on the physical or mental health of a participant, will the participants' Medical Specialist, General Practitioner or Family Doctor be informed of the recruitment of the participant before the research project begins?

(This includes any medical practitioner of whom the participant is a patient).

☐Yes

☐No

☒N/A – No need to inform

Specify the categories of personal data to be collected and analysed: (Select all that apply)

☐Political Opinions

☐Religious or Philosophical Opinions

☐Racial or Ethnic Origin

☐Trade Union Membership

☐Physical Health

☐Mental Health

☐Sex Life

☐Sexual Orientation

☐Alleged Offences or Proven Offences

☐Gender Identity

☒Other (e.g. name, age range, location, interactions, opinions etc.)

Will you be collecting or processing data that might endanger the individual's physical health or safety in the event of a security breach?

☐Yes

☒No

Is the data obtained from external sources publicly available and already truly anonymised at the point you receive the data?

☒Yes

☐No

Will your research involve use of any technology or algorithm which may be perceived as being privacy intrusive, and may have a long lasting or significant effect on individuals? *For example consider any algorithms which may embed bias or discrimination?*

☐Yes

☒No

Will your research involve any systematic monitoring, which would include processes which observe, monitor or control individuals, and may have a long lasting or significant impact on individuals?

☐Yes

☒No

Will your research include creating detailed profiles of individuals which may have a long lasting or significant impact on those individuals?

☐Yes

☒No

Will the project result in you or others making decisions, or taking action against individuals in ways which can have a significant impact on them?

☐Yes

☒No

Specify the data controller(s) for personal data processed in the course of the project - this is normally the University, unless the data is processed under contract to or in partnership with another organisation. Select at least one of the following:

☒Heriot-Watt University

☐Other

State the legal basis for processing the personal data obtained in the course of the project. (Refer to i to determine which of the following to select).

The Data Subject has given consent to the processing of his or her data

Will the project involve transfers of data outside of the above listed countries, to organisations that are not members of Heriot-Watt University Group?

☐Yes

☒No

How long do you intend to retain personally identifiable data, *for example participant's names and contact details, or the participants' unique identifier*?

☒Unless explicit consent is provided for a data subject to be named in the project outputs, personally identifiable data will only be kept for as long as it is necessary to keep the data in order to verify the integrity of the project’s methods and to ensure the validity of the outputs.

☐Other

Are there any ethical issues that you have identified which have not yet been addressed in this application?

☐Yes

☒No

Are there any other documents which may support the ethics application which have not yet been uploaded?

☐Yes

☒No